

PATENT ABSTRACTS OF JAPAN

BC

(11)Publication number : 11-143899

(43)Date of publication of application : 28.05.1999

(51)Int.Cl.

G06F 17/30

G06F 17/27

G06F 17/21

(21)Application number : 09-307726

(71)Applicant : SEIKO EPSON CORP

(22)Date of filing : 10.11.1997

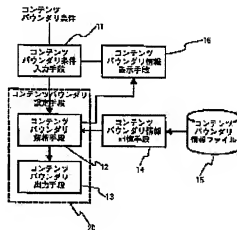
(72)Inventor : TANAKA TOSHIO

(54) DEVICE AND METHOD FOR REGISTER-DOCUMENT PROCESSING AND STORAGE MEDIUM FOR STORING PROGRAM FOR PROCESSING REGISTERED DOCUMENT

(57)Abstract:

PROBLEM TO BE SOLVED: To make it possible to extract contents of an appropriate size in accordance with a processing.

SOLUTION: This processing device which processes a document that has registered contents boundary information for indicating a boundary of unity of document contents, has at least a contents boundary input means 11 that can input contents boundary condition for taking out contents in accordance with the processing, and a contents boundary set means 20 that sets a contents boundary position on the basis of a contents boundary information set for a registered document by contents boundary condition inputted from this contents boundary condition input means 11; and takes out the contents of the registered document on the basis of the contents boundary position information outputted from the contents boundary set means 20.



(19) 日本特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平11-143899

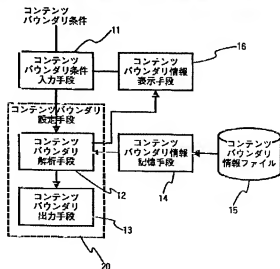
(43) 公開日 平成11年(1999) 5月28日

(51) Int.Cl. ⁴ G 0 6 F 17/30 17/27 17/21	識別記号	F I C 0 6 F 15/40 15/20	3 7 0 A 5 3 0 E 5 7 0 R 5 8 0 L 5 9 0 E
審査請求 未請求 請求項の数 9 O L (全 16 頁) 最終頁に続く			
(21) 出願番号	(71) 出願人 000002369 セイコーエプソン株式会社 東京都新宿区西新宿 2 丁目 4 番 1 号		
(22) 出願日	平成 9 年 (1997) 11 月 10 日		
	(72) 発明者 田中 敏雄 長野県諏訪市大和 3 丁目 3 番 5 号 セイコーエプソン株式会社内		
	(74) 代理人 弁理士 鈴木 喜三郎 (外 2 名)		

(54) 【発明の名称】 登録文書処理装置及び方法並びに登録文書を処理するための処理プログラムを記憶した記憶媒体

(57) 【要約】

【課題】 文書間の差分を取ったり、検索を行ったりというような処理を行う段階でコンテンツを取り出そうとすると、処理量が多く処理に時間を要する問題がある。
【解決手段】 文書内容のほとんどの境界を示すコンテンツバウンダリ情報を有して登録されている文書を処理するためのものであって、処理に応じたコンテンツを取り出すためのコンテンツバウンダリ条件を入力可能なコンテンツバウンダリ入力手段 11 と、このコンテンツバウンダリ条件入力手段 11 から入力されたコンテンツバウンダリ条件により、前記登録文書に対して設定されているコンテンツバウンダリ情報に基づくコンテンツバウンダリ位置を設定するコンテンツバウンダリ設定手段 20 とを少なくとも有し、コンテンツバウンダリ設定手段 20 により出力されたコンテンツバウンダリ位置情報に基づいて登録文書のコンテンツを取り出す。



【特許請求の範囲】

【請求項1】 文書内容のまとまりの境界を示すコンテンツバウンダリ情報が設定されて登録されている文書（登録文書という）を処理するための登録文書処理装置において、
処理に応じたコンテンツを取り出すためのコンテンツバウンダリ条件を入力可能なコンテンツバウンダリ条件入力手段と、
このコンテンツバウンダリ条件入力手段から入力されたコンテンツバウンダリ条件に基づいて、前記登録文書に対して設定されたコンテンツバウンダリ情報を出力し、そのコンテンツバウンダリ情報に対応するコンテンツバウンダリ位置を設定するコンテンツバウンダリ設定手段と、
を少なくとも有することを特徴とする登録文書処理装置。

【請求項2】 前記コンテンツバウンダリ設定手段は、コンテンツバウンダリ条件を受けると、このコンテンツバウンダリ条件と、前記登録文書に対して設定されたそれぞれのコンテンツバウンダリ情報とを比較し、前記コンテンツバウンダリ条件に適合するコンテンツバウンダリ情報を得て、このコンテンツバウンダリ情報に基づくコンテンツバウンダリ位置を設定することを特徴とする請求項1記載の登録文書処理装置。

【請求項3】 文書内容のまとまりの境界を示すコンテンツバウンダリ情報が設定されて登録されている文書（登録文書という）を処理するための登録文書処理装置において、
処理に応じたコンテンツを取り出すために必要なコンテンツバウンダリ条件知識を予め蓄えたコンテンツバウンダリ条件知識ファイルと、
このコンテンツバウンダリ条件知識ファイルの内容に基づいて、前記登録文書に対して設定されたコンテンツバウンダリ情報を処理に応じて出力し、そのコンテンツバウンダリ情報に対応するコンテンツバウンダリ位置を設定するコンテンツバウンダリ設定手段と、
を少なくとも有することを特徴とする登録文書処理装置。

【請求項4】 文書内容のまとまりの境界を示すコンテンツバウンダリ情報が設定されて登録されている文書（登録文書という）を処理するための登録文書処理方法において、
処理に応じたコンテンツを取り出すためのコンテンツバウンダリ条件が入力されると、そのコンテンツバウンダリ条件を受け付け、そのコンテンツバウンダリ条件に基づいて、前記登録文書に対して設定されているコンテンツバウンダリ情報を出力し、このコンテンツバウンダリ情報に対応するコンテンツバウンダリ位置を設定する処理を登録文書処理に含むことを特徴とする登録文書処理方法。

【請求項5】 前記コンテンツバウンダリ条件に基づいて、前記登録文書に対して設定されたコンテンツバウンダリ情報を出力する処理は、入力されたコンテンツバウンダリ条件と、前記登録文書に対して設定されたそれぞれのコンテンツバウンダリ情報とを比較し、前記コンテンツバウンダリ条件に適合するコンテンツバウンダリ情報を得て、このコンテンツバウンダリ情報に基づくコンテンツバウンダリ位置を設定することを特徴とする請求項4記載の登録文書処理方法。

【請求項6】 文書内容のまとまりの境界を示すコンテンツバウンダリ情報が設定されて登録されている文書（登録文書という）を処理するための登録文書処理方法において、
処理に応じたコンテンツを取り出すために必要なコンテンツバウンダリ条件知識を格納したコンテンツバウンダリ条件知識ファイルを有し、このコンテンツバウンダリ条件知識ファイル内容に基づいて、前記登録文書に対して設定されたコンテンツバウンダリ情報を処理に応じて出力し、そのコンテンツバウンダリ情報に対応するコンテンツバウンダリ位置を設定する処理を登録文書処理に含むことを特徴とする登録文書処理方法。

【請求項7】 文書内容のまとまりの境界を示すコンテンツバウンダリ情報が設定されて登録されている文書（登録文書という）を処理するための処理プログラムを記憶した記憶媒体であって、その処理プログラムは、処理に応じたコンテンツを取り出すためのコンテンツバウンダリ条件が入力されると、そのコンテンツバウンダリ条件を受け付け、そのコンテンツバウンダリ条件に基づいて、前記登録文書に対して設定されたコンテンツバウンダリ情報を出力し、このコンテンツバウンダリ情報に対応するコンテンツバウンダリ位置を設定する処理を含むことを特徴とする登録文書を処理するための処理プログラムを記憶した記憶媒体。

【請求項8】 前記コンテンツバウンダリ条件に基づいて、前記登録文書に対して設定されたコンテンツバウンダリ情報を出力する処理は、入力されたコンテンツバウンダリ条件と、前記登録文書に設定されたコンテンツバウンダリ情報とを比較し、前記コンテンツバウンダリ条件に適合するコンテンツバウンダリ情報を得て、このコンテンツバウンダリ情報に基づくコンテンツバウンダリ位置を設定することを特徴とする請求項7記載の登録文書を処理するための処理プログラムを記憶した記憶媒体。

【請求項9】 文書内容のまとまりの境界を示すコンテンツバウンダリ情報が設定されて登録されている文書（登録文書という）を処理するための処理プログラムを記憶した記憶媒体であって、その処理プログラムは、登録文書に対する処理に応じて、処理に応じたコンテンツを取り出すために必要なコンテンツバウンダリ条件をコンテンツバウンダリ条件知識ファイルから取り出し、そ

のコンテンツバウンダリ条件に基づいて、前記登録文書に対して設定されたコンテンツバウンダリ情報を出し、そのコンテンツバウンダリ情報に対応するコンテンツバウンダリ位置を設定する処理を登録文書処理に含むことを特徴とする登録文書処理するための処理プログラムを記憶した記憶媒体。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、コンテンツバウンダリ情報が設定されて登録されている文書（登録文書）を用いて、検索処理や複数の文書間の差分を取ったりする処理を行うための登録文書処理装置及び方法並びに登録文書処理するための処理プログラムを記憶した記憶媒体に関する。

【0002】

【従来の技術】2つの文書の差分を取ったり、ある文書の一部を抽出したり、文書内に書かれていることを検索したりするということに、文書に対しては様々な処理がなされる。

【0003】このような処理を行う場合、文書を段落など、文書の内容のまとまり（コンテンツと呼ぶ）ごとにそのまとまりの境界を示すコンテンツバウンダリを抽出して、そのコンテンツバウンダリにより得られるコンテンツごとに処理を行う方法が従来より用いられている。

【0004】従来では、前述したような2つの文書の差分を取ったりする様々な処理を行う際に、コンテンツバウンダリを抽出して、コンテンツを抽出するのが一般的である。

【0005】なお、検索処理を行う場合は、文書中のキーワードや文字列とその位置をインデックス情報として、文書を登録する時に作成しており、その情報を用いて検索を行うことで検索処理を高速化することが従来より行われている。このような検索処理にあっても、文書をコンテンツに分割する処理は、検索するときに行われるのが普通である。

【0006】このように、従来では、差分を取ったり、検索したりする処理を行う際、これらの処理を行うに必要なコンテンツの抽出は、それらの処理を行うことになるのが普通である。

【0007】しかし、差分を取ったり、検索したりする処理を行う際に、コンテンツを抽出するための処理（コンテンツバウンダリ抽出も含めた処理）を行うと、差分を取ったり、検索したりする処理を行う前に、まず、コンテンツを抽出するための処理を行う必要があるため、処理量が多くなり、処理速度の低下を招くことにもなる。つまり、既に登録されている文書に対し、文書間の差分を取ったり、検索したりする処理を行う場合、これらの処理を行うたびに、その都度、文書の構造解析を行って、コンテンツを抽出する処理を行う必要がある。

【0008】一方、特開平8-272822の「文書登

録装置および文書検索装置」には、文書を登録する際に、文書を所定のブロック（コンテンツと同意のものと考えられるので以下ではコンテンツと表現する）単位に分割し、コンテンツを識別する識別子を付与し、コンテンツ単位からキーワードを抽出し、識別子をキーワードと対応付けしてキーワードのインデックス情報を作成するというような内容が示されている。

【0009】

【発明が解決しようとする課題】前述の特開平8-272822（従来技術という）は、文書を登録する際に、文書を所定のコンテンツ単位に分割して登録することが示されている。しかしながら、この従来技術は、登録時に文書を所定のコンテンツに分割してしまうので、検索などの処理時には、そのコンテンツ単位での処理を行うしかなく、ユーザーの必要とする大きさのコンテンツを取り出したり、処理内容に応じて、コンテンツを適切な大きさに変更することができないという問題がある。

【0010】また、この従来技術におけるコンテンツは、検索を行うためのキーワードが文書中のどこに存在するかを表すために用いられるものであり、文書間の差分を取ったり、特定のコンテンツを抽出したりといった様々なコンテンツ処理を行うためのものではないため、これらの処理には不向きである。

【0011】そこで本発明は、文書を登録する際に、コンテンツの境界を示すコンテンツバウンダリ情報を求め、そのコンテンツバウンダリ情報を有して登録された文書を用いて、検索や文書間の差分を取るといった様々な処理を行い、しかも、既に設定されているコンテンツバウンダリ情報を様々な処理に応じて柔軟に選択可能とすることで、処理に応じた適切な大きさのコンテンツを抽出することを可能とすることを目的としている。

【0012】

【課題を解決するための手段】前述の目的を達成するために、本発明の請求項1に記載された登録文書処理装置の発明は、文書内容のまとまりの境界を示すコンテンツバウンダリ情報が設定されて登録されている文書（登録文書）を処理するための登録文書処理装置において、処理に応じたコンテンツを取り出すためのコンテンツバウンダリ条件を入力可能なコンテンツバウンダリ条件入力手段と、このコンテンツバウンダリ条件入力手段から入力されたコンテンツバウンダリ条件に基づいて、前記登録文書に対して設定されたコンテンツバウンダリ情報を出し、そのコンテンツバウンダリ情報に対応するコンテンツバウンダリ位置を設定するコンテンツバウンダリ設定手段とを少なくとも有した構成としている。

【0013】そして、前記コンテンツバウンダリ設定手段は、コンテンツバウンダリ条件を受けると、このコンテンツバウンダリ条件と、前記登録文書に対して設定されたそれぞれのコンテンツバウンダリ情報とを比較し、前記コンテンツバウンダリ条件に適合するコンテンツバ

ウンダリ情報を得て、このコンテンツバウンダリ情報に基づくコンテンツバウンダリ位置を設定するようにしている。

【0014】また、請求項3に記載された登録文書処理装置の発明は、文書内容のまとりの境界を示すコンテンツバウンダリ情報が設定されて登録されている文書（登録文書）を処理するための登録文書処理装置において、処理に応じたコンテンツを取り出すために必要なコンテンツバウンダリ条件知識を予め蓄えたコンテンツバウンダリ条件知識ファイルと、そのコンテンツバウンダリ条件知識ファイルの内容に基づいて、前記登録文書に対して設定されたコンテンツバウンダリ情報を処理に応じて出力し、そのコンテンツバウンダリ情報に対応するコンテンツバウンダリ位置を設定するコンテンツバウンダリ設定手段とを少なくとも有した構成としている。

【0015】また、本発明の請求項4に記載された登録文書処理方法の発明は、文書内容のまとりの境界を示すコンテンツバウンダリ情報が設定されて登録されている文書（登録文書）を処理するための登録文書処理方法において、処理に応じたコンテンツを取り出すためのコンテンツバウンダリ条件が入力されると、そのコンテンツバウンダリ条件を受け付け、そのコンテンツバウンダリ条件に基づいて、前記登録文書に対して設定されているコンテンツバウンダリ情報を出力し、このコンテンツバウンダリ情報に対応するコンテンツバウンダリ位置を設定する処理を登録文書処理に含むものである。

【0016】そして、前記コンテンツバウンダリ条件に基づいて、前記登録文書に対して設定されたコンテンツバウンダリ情報を出力する処理は、入力されたコンテンツバウンダリ条件と、前記登録文書に対して設定されたそれぞれのコンテンツバウンダリ情報とを比較し、前記コンテンツバウンダリ条件に適合するコンテンツバウンダリ情報を得て、このコンテンツバウンダリ情報に基づくコンテンツバウンダリ位置を設定するようにしている。

【0017】また、本発明の請求項6に記載された登録文書処理方法の発明は、文書内容のまとりの境界を示すコンテンツバウンダリ情報が設定された文書（登録文書）を処理するための登録文書処理方法において、処理に応じたコンテンツを取り出すために必要なコンテンツバウンダリ条件知識を格納したコンテンツバウンダリ条件知識ファイルを用意し、このコンテンツバウンダリ条件知識ファイル内容に基づいて、前記登録文書に対して設定されたコンテンツバウンダリ情報を処理に応じて出力し、そのコンテンツバウンダリ情報に対応するコンテンツバウンダリ位置を設定する処理を登録文書処理に含むことを特徴としている。

【0018】また、請求項7に記載の登録文書処理するための処理プログラムを記憶した記憶媒体は、文書内容のまとりの境界を示すコンテンツバウンダリ情報が

設定されて登録されている文書（登録文書）を処理するための処理プログラムを記憶した記憶媒体であって、その処理プログラムは、処理に応じたコンテンツを取り出すためのコンテンツバウンダリ条件が入力されると、そのコンテンツバウンダリ条件を受け付け、そのコンテンツバウンダリ条件に基づいて、前記登録文書に対して設定されたコンテンツバウンダリ情報を出力し、このコンテンツバウンダリ情報に対応するコンテンツバウンダリ位置を設定する処理を含むものである。

【0019】そして、前記コンテンツバウンダリ条件に基づいて、前記登録文書に対して設定されたコンテンツバウンダリ情報を出力する処理は、入力されたコンテンツバウンダリ条件と、前記登録文書に設定されたコンテンツバウンダリ情報とを比較し、前記コンテンツバウンダリ条件に適合するコンテンツバウンダリ情報を得て、このコンテンツバウンダリ情報に基づくコンテンツバウンダリ位置を設定するようにしている。

【0020】また、本発明の請求項9に記載の登録文書を処理するための処理プログラムを記憶した記憶媒体は、文書内容のまとりの境界を示すコンテンツバウンダリ情報が設定されて登録されている文書（登録文書）を処理するための処理プログラムを記憶した記憶媒体であって、その処理プログラムは、登録文書に対する処理に応じて、処理に応じたコンテンツを取り出すために必要なコンテンツバウンダリ条件をコンテンツバウンダリ知識ファイルから取り出し、そのコンテンツバウンダリ条件に基づいて、前記登録文書に対して設定されたコンテンツバウンダリ情報を出力し、そのコンテンツバウンダリ情報に対応するコンテンツバウンダリ位置を設定する処理を登録文書処理に含むことを特徴としている。

【0021】本発明は、コンテンツバウンダリ情報が設定されて登録された文書を用いて、検索や文書間の差分を取るといった様々な処理を行うものであり、これらの様々な処理を行う際、既に設定されているコンテンツバウンダリ情報を様々な処理に応じて選択することで、コンテンツバウンダリ位置を任意に設定することができる。これにより、処理に応じた適切な大きさのコンテンツを抽出することが可能となる。

【0022】これを実現するための1つの手段として、ユーザがコンテンツバウンダリ条件を入力することにより、そのコンテンツバウンダリ条件に基づいて、前記登録文書に対して設定されているコンテンツバウンダリ情報の中から必要なコンテンツバウンダリ情報を選択し、このコンテンツバウンダリ情報に対応するコンテンツバウンダリ位置を設定する処理を行う。これは、ユーザがコンテンツバウンダリ条件を明示的に入力することによって、処理に必要なコンテンツを取り出すものであり、ユーザの意図を的確に反映したコンテンツを取り出すことができる。

【0023】また、他の手段として、処理に応じたコン

テンツを取り出すために必要なコンテンツバウンダリ条件知識を持ち、このコンテンツバウンダリ条件知識に基づいて、前記登録文書に対して設定されたコンテンツバウンダリ情報の中から、処理に応じて選択して出力し、そのコンテンツバウンダリ情報に対応するコンテンツバウンダリ位置を設定することも可能である。これによれば、処理内容に応じて自動的に、処理に応じたコンテンツ条件が設定され、そのコンテンツ条件に基づいてコンテンツバウンダリ位置が設定されるので、ユーザが殆ど手を加えることなく、自動的に、処理に最適なコンテンツの抽出を行うことができる。

【0024】このように、本発明は、登録文書に細かく設定されたコンテンツバウンダリ情報の中から、コンテンツバウンダリ情報を取捨選択することができるので、登録文書を用いて、文書間の差分を取ったり、検索したりする処理を行う際に、処理に必要なコンテンツを取り出す処理がきわめて簡単に行える。つまり、登録された文書を処理する段階において、予め設定されているコンテンツバウンダリ情報の中から、コンテンツを取り出すためのコンテンツバウンダリ情報を処理の種類に応じて任意に決めることができる。

【0025】

【発明の実施の形態】以下、本発明に実施の形態について説明する。本発明は、登録された文書（登録文書）に対し、検索処理や文書間の差分を取るなど何らかの処理を施す際に文書を文章のまとまりを1つの単位として、そのまとまり（コンテンツ）に分割して取り出すための装置および方法に係わる発明であるが、まず、文書を登録する段階で、登録しようとする文書（以下、登録対象文書という）からコンテンツバウンダリ情報を取得して、そのコンテンツバウンダリ情報を保存する処理についてを説明し、その後で、その登録文書に対して文書を所定のコンテンツにて抽出する処理について説明する。

【0026】図1は文書を登録する際に行われるコンテンツバウンダリ情報取得についての文書登録装置（以下、第1の文書登録装置という）の構成図である。この発明でいう、文書登録装置というのは、文書の作成、編集、保存などが可能な装置であり、たとえば、パーソナルコンピュータなどもその一例である。

【0027】この第1の文書登録装置は、コンテンツバウンダリ入力手段1、コンテンツバウンダリ情報記憶手段2を少なくとも備えた構成となっている。

【0028】このような構成において、登録対象文書3に対して、ユーザが、コンテンツバウンダリ入力手段1からコンテンツバウンダリの位置を指定すると、その位置に対応するコンテンツバウンダリ情報がコンテンツバウンダリ情報記憶手段2に記憶される。このコンテンツバウンダリ情報記憶手段2に記憶内容は、コンテンツバウンダリ情報ファイル4として保存することもできる。そして、そのコンテンツバウンダリ情報は、文書の中に

通常は表示されないデータの形式で埋め込まれて保存されてもよいし、また、コンテンツ情報のみを文書データとは別のデータ（たとえば、コンテンツバウンダリ情報テーブル）として保存されるようにしてもよい。

【0029】このように、第1の文書登録装置では、ユーザが登録対象文書3に対して、明示的にコンテンツバウンダリの位置を決めるものであり、具体的には次のようにして行う。

【0030】たとえば、ディスプレイ画面10上に映し出されている登録対象文書3が図2のような内容であったとする。このような文書内容に対して、ユーザは、その文書内容を見て、マウスなどにより、明示的にコンテンツバウンダリの位置を指定して行く。図2において、矢印はマウスカーソルcを示しており、このマウスカーソルcをユーザの意図する部分に位置させ、その位置でマウスをクリックすることによりコンテンツバウンダリ位置が設定される。図2では設定されたコンテンツバウンダリ位置をb1、b2、b3で示している。なお、このようなコンテンツバウンダリ位置を設定する場合は、システムのアプリケーションをコンテンツバウンダリ設定モードに設定して行う。

【0031】また、この図2の例では、コンテンツバウンダリとする部分にマウスカーソルを位置させてクリックすることでコンテンツバウンダリ位置を指定するようにしたが、これに限らず、たとえば、コンテンツバウンダリで仕切られる文書内容（コンテンツという）の先頭にマウスカーソルを位置させて、そのコンテンツの終わりでマウスカーソルをドラッグさせることによって、コンテンツバウンダリ位置を指定することも可能である。

【0032】このようにして、コンテンツバウンダリ位置の指定がなされるが、そのコンテンツバウンダリ位置に対応するコンテンツバウンダリ情報は、前述したように、文書とは別のデータとして保存してもよく、文書の中に通常は表示されないデータの形式で埋め込んで保存してもよい。

【0033】図3は図2で指定されたコンテンツバウンダリ情報を、文書の中に通常は表示されないデータの形式で埋め込んだ例を示すものである。図3の例では、HTML (Hyper Text Markup Language) のコメントタグを用いて、文書の中に埋め込んだ例である。

【0034】図3において、<!--CB1-->が示される内容がコンテンツバウンダリ情報である。この<!--CB1-->は、「<!--」がコメントの開始を表し、「-->」がコメントの終了を表している。

【0035】一例として、<!--CB1 ファイル端 140 1-->というように記述されたコメントタグが有るとすると、その中のコンテンツバウンダリ情報として、「CB1」はコンテンツバウンダリの識別番号、「ファイル端」はコンテンツバウンダリの種類、「14

0」は、そのコンテンツバウンダリの種類（この場合「ファイル端」）の反対側のコンテンツバウンダリの識別番号であり、そのコンテンツバウンダリ種類により仕切られるコンテンツの大きさも表している。また、「1」はネストレベルを表している。なお、このようなコンテンツバウンダリ情報の詳細については後に説明する。

【0036】また、前述の「CB1 ファイル端 140 1-」で示されるコメントタグには、コンテンツバウンダリの位置を表す内容は存在しないが、そのコンテンツバウンダリの位置は、このようなコメントタグの存在する位置であり、これによってコンテンツバウンダリの位置がわかるのでその位置情報は、特に、記述する必要はない。

【0037】以上が第1の文書登録装置についての説明である。この第1の文書登録装置におけるコンテンツバウンダリ情報の抽出処理は、処理対象文書に対し、ユーザが明示的にコンテンツバウンダリ位置を指示することにより行われるものであり、ユーザが行うべき操作がやや面倒であるが、ユーザの意図する場所に確実にコンテンツバウンダリを設定することができるのが大きな特徴である。

【0038】図4は登録対象文書を登録する際に行われるコンテンツバウンダリ情報取得を行うための第2の文書登録装置の構成図である。

【0039】この第2の文書登録装置は、図1で示した第1の文書登録装置で示されたコンテンツバウンダリ情報記憶手段2、コンテンツバウンダリ情報ファイル4を備え、加えて、コンテンツバウンダリ条件入力手段5とコンテンツバウンダリ解析手段6を備え、さらに、必要に応じて、コンテンツバウンダリ情報表示手段7を備えた構成となっている。

【0040】この第2の文書登録装置では、第1の文書登録装置のように、登録対象文書3に対して、ユーザが、コンテンツバウンダリの位置を明示的に指定するのではなく、コンテンツバウンダリ条件を入力することで、そのコンテンツバウンダリ条件に基づいて、自動的にコンテンツバウンダリを設定する。

【0041】すなわち、コンテンツバウンダリ条件入力手段5により、ユーザがコンテンツバウンダリ条件の入力を行うと、コンテンツバウンダリ解析手段6により、ユーザによって設定されたコンテンツバウンダリ条件を解析する。ここでコンテンツバウンダリ条件というのは、たとえば、段落、空行、改行、リスト、リスト項目、表など文書を1つのまとまりとして抽出できる部分である。このようなコンテンツ条件をユーザが入力すると、コンテンツバウンダリ解析手段6が処理対象文書をたとえばスキャンして、入力されたコンテンツバウンダリ条件に基づいて、コンテンツバウンダリ候補の位置や種類を抽出して、コンテンツバウンダリ情報として出力

する。

【0042】このコンテンツバウンダリ解析手段6によって得られたコンテンツバウンダリ情報は、コンテンツバウンダリ情報記憶手段2に記憶される。このとき、前述の第1の文書登録装置と同様に、このコンテンツバウンダリ情報記憶手段2の記憶内容は、コンテンツバウンダリ情報ファイル4として保存することもできる。そして、そのコンテンツバウンダリ情報は、文書の中に通常は表示されないデータ形式で埋め込まれて保存されてもよいし、また、コンテンツ情報のみを文書データとは別のデータ（たとえば、コンテンツバウンダリ情報テーブル）として保存されるようにしてもよい。

【0043】また、コンテンツバウンダリ解析手段6によって得られたコンテンツバウンダリ情報は、コンテンツバウンダリ情報表示手段7に表示させ、ユーザの設定したコンテンツバウンダリ条件に対してどのようなコンテンツ情報が作成されたかをユーザに知らせるようにすることもできる。

【0044】以下に具体例を参照しながらさらに説明する。

【0045】図5に示すように、たとえば、ディスプレイ画面10に登録対象文書3などの表示を行うための表示エリア10aと、コンテンツバウンダリ条件入力手段としての表示エリア10bとを設定し、表示エリア10aには登録対象文書3を表示し、表示エリア10bにはコンテンツバウンダリ条件を表示する。ここでは、コンテンツバウンダリ条件として、ファイル端、リスト、リスト項目、空行、改行、表などが示されている。

【0046】そして、ユーザがたとえば、コンテンツバウンダリ条件として「リスト」を選択したとする（図5において、選択されたコンテンツバウンダリ条件は黒丸で示されている）。これにより、コンテンツバウンダリ解析手段6は、ユーザの設定したコンテンツバウンダリ条件に基づいて、登録対象文書3内のコンテンツバウンダリとなりうるコンテンツバウンダリ候補の位置および種類の解析を行い、その結果をコンテンツバウンダリ情報として出力する。

【0047】そして、コンテンツバウンダリ解析手段6により得られるコンテンツバウンダリ情報に基づいて、コンテンツバウンダリ位置をディスプレイ画面10の表示エリア10a上で表示するとともに、前述の第1の文書登録装置で説明したように、コンテンツバウンダリ解析手段6により得られるコンテンツバウンダリ情報をコンテンツバウンダリ情報記憶手段2に記憶させる。

【0048】ユーザはディスプレイ画面10の表示エリア10aに表示された内容を見て、自分の意図したコンテンツバウンダリ位置が適正に反映されているか否かを判断し、修正したい箇所があればそれを指示することもできる。

【0049】なお、コンテンツバウンダリ条件は、図5

で示した例のように、予め表示されている度つかの条件から選択するというのではなく、ユーザがコンテンツバウンダリ条件や、コンテンツバウンダリとなりうるパターンなどを入力するようにしてもよい。このコンテンツバウンダリとなりうるパターンというのは、たとえば、文書中に、規則性のある字句や記号が繰り返し現れるような場合、その字句や記号をコンテンツバウンダリとして入力することもできる。たとえば、具体例として、「1日」という項目があって、その「1日」という項目のあとに、あるまとまった文章が存在し、行を変えて、「2日」という項目があって、その「2日」という項目についてあるまとまった文章が存在するというように、規則性のある字句や記号が繰り返し現れるような文書内容があるとする。このような例では、「数字+日」といったパターンをコンテンツバウンダリ条件として設定することができる。

【0050】以上説明した第2の文書登録装置では、ユーザがコンテンツバウンダリ条件を入力するだけで、あとは、入力されたコンテンツバウンダリ条件に基づいて、自動的に処理対象文書に対するコンテンツバウンダリ情報を得ることができる。なお、この第2の文書登録装置は、第1の文書登録装置と同様に、ユーザが処理対象文書中に明示的にコンテンツバウンダリを指示することも可能である。

【0051】図6は登録対象文書を登録の際に行われるコンテンツバウンダリ情報取得を行うための第3の文書登録装置の構成図である。

【0052】この第3の文書登録装置は、図4の第2の文書登録装置で示されたコンテンツバウンダリ情報記憶手段2とコンテンツバウンダリ解析手段6、コンテンツバウンダリ情報ファイル4を備え、加えて、コンテンツバウンダリ知識が格納されたコンテンツバウンダリ知識ファイル8を備えた構成となっている。

【0053】この第3の文書登録装置では、登録対象文書3に対し、コンテンツバウンダリ知識ファイル8を用いて、コンテンツバウンダリとなりうる部分をすべて自動的に検出し、それをコンテンツバウンダリ情報としてコンテンツバウンダリ情報記憶手段2に記憶させるものである。

【0054】前述のコンテンツバウンダリ知識ファイル8には、コンテンツバウンダリとなりうる各種の条件が予め記述されている。このコンテンツバウンダリとなりうる各種の条件というのは、たとえば、句点、改行、空白、大文字見出し、HTMLにおけるタグなどである。

【0055】図7は第3の文書登録装置の処理手順を説明するフローチャートであり、まず、登録対象文書を入力して(ステップs1)、データの読み込みを行い、文書末であるか否かを判定し(ステップs2)、文書末であれば終了し、文書末でなければステップs3に進む。

ステップs3では、登録対象文書にコンテンツバウンダリ候補が存在するか否かを判定し、存在しなければ、ステップs2に戻り、コンテンツバウンダリ候補が有れば、そのコンテンツバウンダリ候補に対してコンテンツバウンダリ情報を設定し、そのコンテンツバウンダリ情報をコンテンツバウンダリ情報記憶手段に記憶させる(ステップs4)。

【0056】この図7のフローチャートで示される処理は、主に、コンテンツバウンダリ解析手段6が行う処理であり、読み込んだ登録対象文書と、コンテンツバウンダリ知識ファイルに記述されているコンテンツバウンダリ条件とを比較し、登録対象文書内にコンテンツバウンダリ候補を検出したとして、その部分に対応するコンテンツバウンダリ情報を得て、そのコンテンツバウンダリ情報をコンテンツバウンダリ記憶手段2に記憶させる。

【0057】そして、コンテンツバウンダリ記憶手段2では、受け取ったコンテンツバウンダリ情報と新たなコンテンツバウンダリ情報として格納する。このとき、前述の第1、第2の文書登録装置と同様に、コンテンツバウンダリ情報記憶手段2の記憶内容は、コンテンツバウンダリ情報ファイル4として保存することもできる。そして、そのコンテンツバウンダリ情報は、文書の中に通常は表示されないデータの形式で埋め込まれて保存されてもよいし、また、コンテンツバウンダリ情報のみを文書データとは別のデータ(たとえば、コンテンツバウンダリ情報テーブル)として保存されるようにしてもよい。

【0058】以上説明した第3の文書登録装置では、登録対象文書に対し、コンテンツバウンダリ知識ファイル8の内容に基づいて、自動的にコンテンツバウンダリ情報を得ることができ、ユーザがコンテンツバウンダリを明示的に指示したり、コンテンツバウンダリ条件を入力したりする操作が不要となる。

【0059】ところで、これまで説明した第1～第3の文書登録装置にて求められるコンテンツバウンダリ情報は、たとえば、図8に示すようなコンテンツバウンダリ情報テーブルとして表すことができる。以下、このコンテンツバウンダリ情報について図8のコンテンツバウンダリ情報テーブルを参照しながら説明する。

【0060】図8で示されるコンテンツバウンダリ情報テーブルは、そのテーブルの1つの行で示される内容が1つのコンテンツバウンダリ情報であり、たとえば、第1行目の内容、つまり、コンテンツバウンダリ識別番号「1」、コンテンツバウンダリ位置「0」、コンテンツバウンダリの種類「ファイル端」、対応するコンテンツバウンダリ「140」、ネストレベル「1」は、図3において、HTMLのコメントタグの一例として示した<!--CB1 ファイル端 140 1-->に対応するコ

ンテンツバウンダリ情報である。

【0061】このようなコンテンツバウンダリ情報において、コンテンツバウンダリ識別番号は、その文書のコンテンツバウンダリとなりうる部分に付された番号である。

【0062】また、コンテンツバウンダリ位置は、文書データの先頭からの文字数を表し、コンテンツバウンダリ位置が「0」というのは、文書ファイルの先頭を表している。また、コンテンツバウンダリの種類は、コンテンツバウンダリが文書のどのような部分かを表すもので、コンテンツバウンダリの種類が「ファイル端」というのは、コンテンツバウンダリがその文書の端であることを表している。

【0063】そして、対応するコンテンツバウンダリというのは、コンテンツバウンダリの種類で指定されるコンテンツバウンダリの反対側に位置するコンテンツバウンダリの位置を、コンテンツバウンダリ識別番号で表すものである。

【0064】たとえば、コンテンツバウンダリの種類が「ファイル端」で対応するコンテンツバウンダリが「140」であるとする、ファイルの先頭の反対側のコンテンツバウンダリ位置、つまり、ファイルの終わりの位置が、コンテンツバウンダリ識別番号「140」であることを表している。

【0065】また、図3のような文書内容において、HTMLのコメントタグが、<!--CB15 リスト項目 15 3-->となっている場合は、図8で示されるコンテンツバウンダリ情報テーブルからわかるように、コンテンツバウンダリ識別番号は「15」であり、コンテンツバウンダリ位置の「50」は、文書のファイル先頭からの文字数が51文字目（先頭が0から始まっているので、「50」は51文字目となる）を表している。

【0066】また、コンテンツバウンダリの種類が「リスト項目」というのは、リストとして記述された幾つかの項目のうち1つの項目であることを表している。そして、対応するコンテンツバウンダリ「15」は、この場合、そのリスト項目自体を1つのコンテンツとすることを意味している。

【0067】また、ネストレベルを示す数値は、このコンテンツバウンダリ情報テーブルで表されるように、最も大きなコンテンツをその文書ファイル全体としたとき、その文書ファイル全体のネストレベルを「1」とし、その中に、たとえば、リストという内容が1つのコンテンツとして存在した場合、そのリストによるコンテンツは、その文書ファイル全体で表されるコンテンツの中に含まれるので、ネストレベルを「2」とし、そのリストの中に存在するリスト項目は、ネストレベルを「3」とするというように、あるコンテンツの中に含まれるコンテンツ、さらにそのコンテンツの中に含まれるコンテンツというように、包含される度合いが高いほど

ネストレベルを表す数値が大きくなる。

【0068】また、図8のコンテンツバウンダリ情報テーブルにおいて、たとえば、コンテンツバウンダリ識別番号「3」のコンテンツバウンダリは、そのコンテンツバウンダリ位置が「30」であり、コンテンツバウンダリの種類が「句点」で、対応するコンテンツバウンダリが「2」、ネストレベルが「2」となっている。これは、対応するコンテンツバウンダリが「2」であることから、この場合、「句点」でコンテンツを仕切ると、ファイルの11文字目から31文字目（先頭が0から始まっているので、「10」は11文字目、「30」は31文字目となる）までを1つのコンテンツとするということであり、そのネストレベルは、ファイル全体を1つのコンテンツとして考えたとき、そのコンテンツ内に含まれるため、ネストレベルが「2」となっている。

【0069】以上のようにして、処理対象文書のコンテンツバウンダリ情報が作成され、そのコンテンツバウンダリ情報がコンテンツバウンダリ情報記憶手段2に記憶される。この図8に示すコンテンツバウンダリ情報テーブルにおいては、そのコンテンツバウンダリ情報テーブルにおける1つの行がそれぞれのコンテンツバウンダリ情報を示している。

【0070】なお、このようなコンテンツバウンダリ情報において、バウンダリの種類は、コード化してもよい。たとえば、「ファイル端」は「1」、「句点」は「2」、「リスト」は「3」というようにコード化して、そのコードデータを記憶するようにしてもよい。また、バウンダリ位置は文字数でなくとも、バイト数でもよく、また行数でもよい。

【0071】以上が文書を登録する際には、コンテンツバウンダリ情報を得て文書の登録を行う処理である。ところで、これまでの説明では、登録対象文書そのものを登録する処理についての説明はなされていないが、この登録対象文書は所定の登録手段に登録されることはいうまでもない。この登録は、前述したように、コンテンツバウンダリ情報とは別のデータとして登録されてもよく、あるいは、コンテンツバウンダリ情報が埋め込まれた状態で登録されてもよい。

【0072】このようにして、コンテンツバウンダリ情報を持って登録された文書に対し、検索処理や複数の文書間の差分をとるなど何らかの処理を行うとする際、すでに設定されたコンテンツバウンダリ情報を用いることで、これらの様々な処理に対応した適切なコンテンツを取り出すことができ、それぞれの処理を円滑にかつ容易に行うことができる。なお、処理の種類などによっては、その処理の種類に応じた適切な大きさのコンテンツを抽出する必要がある。このように、登録時においては、コンテンツそのものを決めるのではなく、コンテンツを取り出すためのコンテンツバウンダリ情報を細かく求めているので、登録された文書を処理する段階で、コ

ンテンツパウングリ情報を取捨選択することが可能であり、それによって、取り出すコンテンツの大きさを柔軟に設定することができ、処理に対応したコンテンツを取り出すことを可能としている。

【0073】なお、これまで説明した第1～第3の文書登録装置では、登録対象文書のコンテンツパウングリ情報を得て、そのコンテンツパウングリ情報をコンテンツパウングリ情報テーブルとして保存したり、コンテンツ情報を文書中に埋め込んで保存したりすることを可能としている。

【0074】したがって、このように登録された文書（登録文書）は、コンテンツ解析が行われているので、その登録文書に対して検索処理を行ったり、文書間の差分を取ったりする処理を行おうとする際、すでに設定されたコンテンツパウングリ情報を用いることで、様々な処理に対応できるが、処理の内容などによっては、処理内容に応じた適切なコンテンツパウングリ位置を設定する必要がある場合もある。これに対処するために、登録文書に対して、以下のようなコンテンツ抽出処理を行う。

【0075】図9は前述したような登録文書に対し適切な大きさのコンテンツを抽出する処理を行うための登録文書処理装置（以下、第1の登録文書処理装置という）の構成図である。

【0076】この第1の登録文書処理装置は、コンテンツパウングリ条件入力手段11、コンテンツパウングリ解析手段12、コンテンツパウングリ出力手段13、コンテンツパウングリ情報記憶手段14、コンテンツパウングリ情報ファイル15、コンテンツパウングリ情報表示手段16を備えている。

【0077】そして、コンテンツパウングリ解析手段12と、コンテンツパウングリ出力手段13とによって、コンテンツパウングリ設定手段20を構成している。このコンテンツパウングリ設定手段20は、コンテンツパウングリ条件が入力されると、そのコンテンツパウングリ条件に基づいて、前記文書登録時に得られたコンテンツパウングリ情報を得て、このコンテンツパウングリ情報によりコンテンツパウングリ位置を示す情報を出して、文書に対し、コンテンツパウングリを設定するものである。

【0078】なお、文書登録と登録文書処理を1つの同じシステムで行う場合は、コンテンツパウングリ条件入力手段11、コンテンツパウングリ解析手段12、コンテンツパウングリ情報記憶手段14、コンテンツパウングリ情報ファイル15、コンテンツパウングリ情報表示手段16などは、前述した文書登録装置（たとえば、第2の文書登録装置）で示したコンテンツパウングリ条件入力手段5、コンテンツパウングリ解析手段6、コンテンツパウングリ情報記憶手段2、コンテンツパウングリ情報ファイル4、コンテンツパウングリ情報表示手段7

と共用することができるが、ここでは、説明の都合上、これらを前述の文書登録装置とは別な符号を付して説明する。

【0079】さらに、これらの構成要素の他に、コンテンツパウングリ設定手段20により設定されたコンテンツパウングリ位置によって、文書をコンテンツに分割してそのコンテンツを抽出する手段なども実際には設けられるが、図9ではこれらを省略している。

【0080】ところで、コンテンツパウングリ情報記憶手段14には、前述の文書登録装置によって得られたコンテンツパウングリ情報が格納されている。また、このコンテンツパウングリ情報は、コンテンツパウングリ情報ファイル15にファイルとして格納されている。なお、そのコンテンツパウングリ情報は、前述したように、文書データとは別のデータとして登録されているように、あるいは、文書データの中に埋め込まれた状態で登録されている。また、

【0081】このような構成において、前述のようにして登録された文書を用いて何らかの処理を行うとする場合、その処理に必要なコンテンツを取り出すためのコンテンツパウングリ条件をコンテンツパウングリ条件入力手段11より、ユーザが入力する。

【0082】また、ユーザが、コンテンツパウングリ条件を入力する。このコンテンツパウングリ条件の入力は、明示的な条件（たとえば、改行、リストなど）を入力してもよいし、ユーザが文書中のコンテンツパウングリ位置を指定し、その指定した位置に存在するコンテンツパウングリ情報から、システムが最適なものを選択するようにしてもよい。さらに、「数字+日」といったパターンをコンテンツパウングリ条件とすることもできる。なお、入力されたコンテンツパウングリ条件が文書登録時にコンテンツパウングリ情報に含まれていない場合には、入力されたコンテンツパウングリ条件を解析することと、それを新たに追加することも可能である。

【0083】以下、この第1の文書処理装置の動作例について説明する。

【0084】たとえば、図8のようなコンテンツパウングリ情報テーブルが得られている場合、ユーザがコンテンツパウングリ位置「43」を指定すれば、「リスト」がコンテンツパウングリ条件となり、コンテンツパウングリ位置の「43」～「89」までの内容を1つのコンテンツとする。

【0085】図10（a）は「改行」と「リスト項目」をコンテンツパウングリ条件とした場合、図10（b）は「リスト項目」をコンテンツパウングリ条件とした場合のコンテンツパウングリ候補位置をそれぞれ示すもので、図においてマークMがコンテンツパウングリ位置を示している。このように、ディスプレイ画面10上の表示エリア10bに表示されるコンテンツパウングリ条件の選択の仕方によって表示エリア10aのコンテンツ

バウンダリ位置が変化する。

【0086】コンテンツツバウンダリ解析手段12は、ユーザによって指定されたコンテンツツバウンダリ条件に適合するコンテンツツバウンダリ情報を、コンテンツツバウンダリ情報記憶手段14から抽出して出力する。

【0087】たとえば、コンテンツツバウンダリ情報が、図8で示されるようなコンテンツツバウンダリ情報テーブルの形式で保存されている場合、ユーザの指定したコンテンツツバウンダリ条件が「リスト項目」であるとする、バウンダリ識別番号「15」、「17」が抽出されるが、このとき、文書の基本構造を作っている識別番号「1」と「140」で表されるファイル端、リスト構造を作っている識別番号「13」と「24」で表されるリストも抽出する。

【0088】これにより、抽出されるコンテンツとしては、そのコンテンツをコンテンツツバウンダリ位置で表すと、「0」～「42」、つまり、ファイル端（ファイルの先頭）からリスト（リスト開始）の前までが1つのコンテンツとして抽出され、コンテンツツバウンダリ位置「50」～「56」、つまり、リスト項目（最初のリスト項目）から2番目のリスト項目の前までが1つのコンテンツとして抽出され、コンテンツツバウンダリ位置「57」～「・・・」、つまり、2番目のリスト項目からそのリスト項目の終わりまで（この表では位置が記載されていないので、「・・・」で表す）が1つのコンテンツとして抽出され、コンテンツツバウンダリ位置「・・・」～「88」、つまり、あるリスト項目の終わりからリスト（リストの終わり）の前までが1つのコンテンツとして抽出されるように、文書の内容のまとまりがそれぞれコンテンツとして抽出されることになる。

【0089】図11はコンテンツツバウンダリ情報がコンテンツツバウンダリ情報テーブル形式である場合の処理の手順を示すフローチャートである。この図11に示す処理手順は、主に、コンテンツツバウンダリ設定手段20が行う処理であり、以下、このフローチャートを参照しながら説明する。

【0090】まず、コンテンツツバウンダリ情報テーブルの内容を読み込む（ステップs11）。そして、ユーザによってコンテンツツバウンダリ条件が入力されると、そのコンテンツツバウンダリ条件を受け付ける（ステップs12）。次に、読み込んだコンテンツツバウンダリ情報のうち1つのコンテンツツバウンダリ情報（コンテンツツバウンダリ情報テーブルの1行分のコンテンツツバウンダリ情報）を読み込む。このとき、読み込むべきコンテンツツバウンダリ情報が存在するか否かを判定し（ステップs13）、コンテンツツバウンダリ情報が存在すれば、その読み込んだコンテンツツバウンダリ情報が、ユーザからによって与えられたコンテンツツバウンダリ条件に適合するか否かを判定する（ステップs14）。そして、読み込んだコンテンツツバウンダリ情報がコンテンツツバウンダリ条

件に適合すれば、そのコンテンツツバウンダリ情報を出力する（ステップs15）。

【0091】次に、再びステップs13に戻って、読み込んだコンテンツツバウンダリ情報のうち次のコンテンツツバウンダリ情報（コンテンツツバウンダリ情報テーブルにおける次の1行分のコンテンツツバウンダリ情報）を読み込み、読み込むべきコンテンツツバウンダリ情報が存在するか否かを判定し（ステップs13）、コンテンツツバウンダリ情報が存在すれば、その読み込んだコンテンツツバウンダリ情報がユーザによって与えられたコンテンツツバウンダリ条件に適合するか否かを判定する（ステップs14）。そして、読み込んだコンテンツツバウンダリ情報がコンテンツツバウンダリ条件に適合すれば、そのコンテンツツバウンダリ情報を出力する（ステップs15）。

【0092】このような処理をコンテンツツバウンダリ情報テーブルのすべてのコンテンツツバウンダリ情報について行い、コンテンツツバウンダリ情報テーブルのすべてのコンテンツツバウンダリ情報について処理が終了すると、これまでの処理により選択されたコンテンツツバウンダリ情報に対応するコンテンツツバウンダリ位置を示すマークが付された文書をコンテンツツバウンダリ情報表示手段16に表示する（ステップs16）。

【0093】ここで、ユーザがその表示内容（たとえば、図10）を見て、ユーザの意図する位置にコンテンツツバウンダリが適切に付されているか否かを判定（ステップs17）、適切に付加されていれば、OKとして処理を終了し、ユーザの意図しない位置にコンテンツツバウンダリが付加されているような場合、あるいは、ユーザの意図する位置にコンテンツツバウンダリが付されていない場合は、コンテンツツバウンダリ条件を変更するなどして、再度、ステップs12以降の処理を行う。

【0094】このように、ユーザがコンテンツツバウンダリ条件を与えることによって、そのコンテンツツバウンダリ条件に適合するコンテンツツバウンダリ情報を選んで、そのコンテンツツバウンダリ位置を文書中に設定することができ、しかも、そのコンテンツツバウンダリ位置をコンテンツツバウンダリ情報表示手段16によってディスプレイ画面上に表示させることにより、ユーザは、その表示内容を見て、コンテンツツバウンダリ位置が適切に設定されているか否かを判断することができ、適切であれば、コンテンツツバウンダリ条件を設定し直すなどして、再度、コンテンツツバウンダリ位置の設定処理を行うことも可能となる。また、ディスプレイ画面上に表示された文書に対してマウスなどで直接、コンテンツツバウンダリ位置を指示することもできる。

【0095】以上の処理は、コンテンツツバウンダリ情報が、コンテンツツバウンダリ情報テーブル形式で保存されている場合の処理手順を示すフローチャートである。これに対して、コンテンツツバウンダリ情報が、文書中にたとえばHTMLなどのコメントタグとして埋め込まれて

いる場合の処理は、図12のような処理手順で行われる。以下、図12のフローチャートを参照しながらその処理手順について説明する。

【0096】まず、ユーザによってコンテンツバンダリ条件の入力が行われると、そのコンテンツバンダリ条件を受け付ける(ステップs21)。そして、文書データ(たとえば、HTMLなどのコメントタグによるコンテンツバンダリ情報が埋め込まれている文書データ)を読み込む。このとき、読み込むべき文書データが存在するか否かを判定し(ステップs22)、文書データが存在すれば、その文書データ中に存在する最初のコメントタグで示されるコンテンツバンダリ情報が、ユーザによって与えられたコンテンツバンダリ条件に適合するか否かを判定する(ステップs23)。そして、埋め込まれているコンテンツバンダリ情報がコンテンツバンダリ条件に適合すれば、そのコンテンツバンダリ情報を出力する(ステップs24)。

【0097】次に、再びステップs22に戻って、読み込むべき文書データが存在するか否かを判定し、文書データがあれば、その文書データ中に存在するコメントタグで示されるコンテンツバンダリ情報のうち、2番目のコンテンツバンダリ情報が、ユーザによって与えられたコンテンツバンダリ条件に適合するか否かを判定する(ステップs23)。そして、読み込んだコンテンツバンダリ情報が入力されたコンテンツバンダリ条件に適合すれば、そのコンテンツバンダリ情報を出力する(ステップs24)。

【0098】このような処理を文書データの終わりまで行い、処理が終了すると、これまでの処理により選択されたコンテンツバンダリ情報に対応するコンテンツバンダリ位置を示すマークの付された文書をコンテンツバンダリ情報表示手段16によってディスプレイ画面上に表示する(ステップs25)。

【0099】ここで、ユーザがその表示内容(たとえば、図10)を見て、ユーザの意図する位置にコンテンツバンダリが適切に付されているか否かを判定し(ステップs26)、ユーザの意図する位置にコンテンツバンダリが適切に付加されていれば、OKとして処理を終了し、ユーザの意図しない位置にコンテンツバンダリが付加されているような場合、あるいは、ユーザの意図する位置にコンテンツバンダリが付されていない場合は、コンテンツバンダリ条件を変更するなどして、再度、ステップs21以降の処理を行う。

【0100】この場合も前述したと同様に、ユーザがコンテンツバンダリ条件を与えることによって、そのコンテンツバンダリ条件に適合するコンテンツバンダリ情報を選んで、そのコンテンツバンダリ位置を文書中に設定することができる。また、そのコンテンツバンダリ位置が設定された文書をコンテンツバンダリ情報表示手段16によってディスプレイ画面上に表示させ

ることにより、ユーザは、その表示内容を見て、コンテンツバンダリ位置が適切に設定されているか否かを判断することができ、適切でなければ、コンテンツバンダリ条件を設定し直すなどして、再度、コンテンツバンダリ位置の設定処理を行うことも可能となる。また、ディスプレイ画面上に表示された文書に対してマウスなどで直接、コンテンツバンダリ位置を指示することもできる。

【0101】以上のような処理を行うことによって、処理対象文書に対し、コンテンツバンダリ位置の設定が任意に行え、設定されたコンテンツバンダリ位置によって、コンテンツを抽出することができる。これにより、検索処理や文書間の差分を取るなどといった様々な処理に応じた適切なコンテンツを取り出すことができる。

【0102】これら図11および図12で説明した2つの処理において、図11の処理は、コンテンツバンダリ情報テーブルにおけるコンテンツバンダリ識別番号の数だけ処理を繰り返せばよく、一方、図12の方は、文書データの文字数分の処理を繰り返す必要がある。たとえば、前述の図8のコンテンツバンダリ情報テーブルの例で考えると、図11の処理は、コンテンツバンダリ情報の数は140個であるため、140回処理を繰り返せばよいが、図12の方は、文字数が0~6408の6409個存在するため、6409回処理を繰り返す必要があるため、図11の方が処理量が少なくて済む。

【0103】図13は前述したような登録文書に対し適切な大きさのコンテンツを抽出する処理を行うための第2の登録文書処理装置の構成図である。

【0104】この第2の登録文書処理装置は、コンテンツバンダリ出力手段13とコンテンツバンダリ情報検索手段17で構成されるコンテンツバンダリ設定手段20と、コンテンツバンダリ条件知識ファイル21を備えた構成となっている。

【0105】そして、この第2の登録文書処理装置は、ユーザがコンテンツバンダリ条件を設定するのではなく、処理の内容に応じて自動的にコンテンツバンダリ条件の設定を可能としたものであり、それを実現するために、コンテンツバンダリ条件知識ファイル21を備えている。このように、この第2の登録文書処理装置では、コンテンツバンダリ条件知識ファイル21によって、自動的に、処理に最適なコンテンツ条件を得て、そのコンテンツ条件に基づいて、処理対象の文書(登録文書22)に対してコンテンツバンダリ位置を設定することが可能となるので、ユーザがコンテンツ条件を入力するためのコンテンツバンダリ条件入力手段や、ユーザの入力したコンテンツバンダリ条件に基づいたコンテンツバンダリ情報を表示するためのコンテンツバンダリ情報表示手段などは特に必要ではない。

【0106】コンテンツバンダリ条件知識ファイル2

1は、処理内容に応じて最適なコンテンツバンダリ条件が蓄えられているもので、たとえば、この第2の文書処理装置を情報検索システムに用いるものとする、情報検索に最適なコンテンツバンダリ条件についての知識が蓄えられている。

【0107】なお、コンテンツバンダリ条件知識ファイル21は、そのシステムが行う処理に合わせたコンテンツバンダリ条件をのみを持たせることも可能であるが、様々な処理内容に適応できるように、様々な処理内容に応じた最適なコンテンツバンダリ条件知識を蓄えておくことも勿論可能である。この場合、たとえば、ユーザが処理の種類などについての指令を、コンテンツバンダリ情報検索手段17に与えることで、コンテンツバンダリ情報検索手段17は、その指令に基づいて、処理に最適なコンテンツバンダリ条件を、コンテンツバンダリ条件知識ファイル21から得て、その条件に基づいて、登録文書22（コンテンツバンダリ情報が埋め込まれている文書）を読み出して、与えられたコンテンツバンダリ条件に対応するコンテンツバンダリ情報を自動的に検索する。

【0108】このように、第2の登録文書処理装置は、システム側の持っているコンテンツバンダリ条件知識によつて、文書の中に埋め込まれているコンテンツバンダリ情報の中から自動的に、コンテンツバンダリ条件知識ファイル21が持っているコンテンツバンダリ条件に適合するコンテンツバンダリ情報を検索するので、処理対象文書に対し、より一層、自動化されたコンテンツバンダリ位置設定が可能となる。

【0109】以上説明したように、本発明では、登録された文書を、たとえば、検索処理や文書間の差分を取るなど様々な処理に用いる場合、登録時に設定されたコンテンツバンダリ情報を用いてコンテンツバンダリ位置を決め、これにより、処理に必要なコンテンツの抽出を行うようにしているので、コンテンツの取り出しがきわめて簡単に行うことができる。たとえば、処理の内容に応じて、文書登録の際に設定された多数のコンテンツバンダリ情報のうち、必要なコンテンツバンダリ情報のみ選択することが可能であり、これにより、コンテンツバンダリ位置を任意に決めることができ、処理の内容に応じて、ユーザの意図する大きさのコンテンツの抽出が可能となる。また、文書登録のときに設定されなかったコンテンツバンダリ情報を、文書処理の時に新たに追加するということが比較的容易に行うことが可能となる。

【0110】なお、本発明は以上説明した各実施の形態に限定されるものではなく、本発明の要旨を逸脱しない範囲で種々変形実施可能となるものである。また、以上説明した本発明の登録文書の処理を行うための処理プログラムは、フロッピーディスク、光ディスク、ハードディスクなどの記録媒体に記録しておくことができ、本

発明はその記録媒体をも含むものである。また、ネットワークから処理プログラムを得るようにしてもよい。

【0111】

【発明の効果】本発明では、コンテンツバンダリ情報が設定された登録文書を、たとえば、検索に用いたり、文書間の差分を取るなど様々な処理に用いる場合、ユーザがコンテンツバンダリ条件を入力することにより、そのコンテンツバンダリ条件に基づいて、前記登録文書に対して設定されているコンテンツバンダリ情報の中から必要なコンテンツバンダリ情報を選択し、このコンテンツバンダリ情報に対応するコンテンツバンダリ位置を設定することができる。これは、ユーザがコンテンツバンダリ条件を明示的に入力することによって、処理に必要なコンテンツを取り出すものであり、ユーザの意図を的確に反映したコンテンツを取り出すことができる。

【0112】また、コンテンツバンダリ条件を入力することなく、自動的に、処理に最適なコンテンツの抽出を行うことができる。これは、処理に応じたコンテンツを取り出すために必要なコンテンツバンダリ条件知識を持ち、このコンテンツバンダリ条件知識に基づいて、前記登録文書に対して設定されたコンテンツバンダリ情報の中から、処理に応じて選択して出力し、そのコンテンツバンダリ情報に対応するコンテンツバンダリ位置を設定するものであり、これによれば、処理内容に応じて自動的に、処理に応じたコンテンツ条件が設定され、そのコンテンツ条件に基づいてコンテンツバンダリ位置が設定されるので、ユーザが殆ど手を加えることなく、自動的に、処理に最適なコンテンツの抽出を行うことができる。

【0113】このように、本発明では、登録時に得られたコンテンツバンダリ情報をもそのまま用いるのではなく、文書を処理する段階で、コンテンツバンダリ情報を取捨選択することが可能であつて、これによって、任意の位置にコンテンツバンダリの設定が可能となり、取り出すコンテンツの大きさを柔軟に設定することができ、処理に対応したコンテンツを取り出すことができる。このように、コンテンツの大きさを柔軟に決めることができることから、検索処理や文書間の差分を取るなどの処理以外にも、文書を所定のコンテンツに分割して処理を行う必要のある様々な処理に対応することができる。

【図面の簡単な説明】

【図1】本発明の登録文書処理装置を実現するために必要な文書登録装置の第1の構成例（第1の文書登録装置）を説明するブロック図。

【図2】第1の文書登録装置におけるコンテンツバンダリ位置指定についての一例を説明する図。

【図3】第1の文書登録装置におけるコンテンツバンダリ情報を文書中に埋め込んだ例を示す図。

【図4】本発明の登録文書処理装置を実現するために必要な文書登録装置の第2の構成例（第2の文書登録装置）を説明するブロック図。

【図5】第2の文書登録装置におけるコンテンツバウンダリ条件入力を行う例を説明する図。

【図6】本発明の登録文書処理装置を実現するために必要な文書登録装置の第3の構成例（第3の文書登録装置）を説明するブロック図。

【図7】第3の文書登録装置におけるコンテンツバウンダリ情報を抽出する処理を説明するフローチャート。

【図8】第1～第3の文書登録装置において得られるコンテンツバウンダリ情報をコンテンツバウンダリ情報テーブルとして表した図。

【図9】本発明の実施の形態である第1の登録文書処理装置を説明するブロック図。

【図10】本発明の実施の形態においてコンテンツバウンダリ条件の設定の仕方の違いによるコンテンツバウンダリ位置の変化を説明する図。

【図11】第1の登録文書処理装置においてコンテンツバウンダリ情報がコンテンツバウンダリ情報テーブルで

ある場合の処理手順を説明するフローチャート。

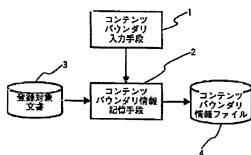
【図12】第1の登録文書処理装置においてコンテンツバウンダリ情報が文書中に埋め込まれた場合の処理手順を説明するフローチャート。

【図13】本発明の実施の形態である第2の登録文書処理装置を説明するブロック図。

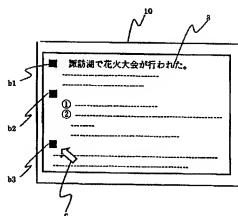
【符号の説明】

- 1 コンテンツバウンダリ入力手段
- 2, 14 コンテンツバウンダリ情報記憶手段
- 3 登録対象文書
- 4, 15 コンテンツバウンダリ情報ファイル
- 5, 11 コンテンツバウンダリ条件入力手段
- 6, 12 コンテンツバウンダリ解析手段
- 7, 16 コンテンツバウンダリ情報表示手段
- 8 コンテンツバウンダリ知識ファイル
- 13 コンテンツバウンダリ出力手段
- 17 コンテンツバウンダリ情報検索手段
- 20 コンテンツバウンダリ設定手段
- 21 コンテンツバウンダリ条件知識ファイル
- 22 登録文書

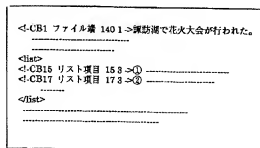
【図1】



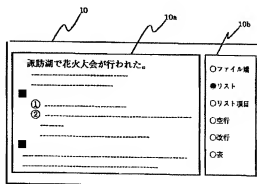
【図2】



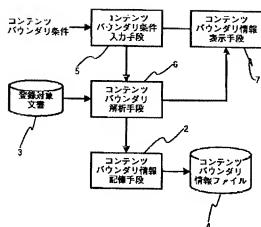
【図3】



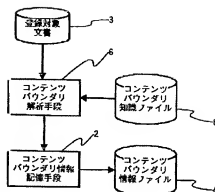
【図5】



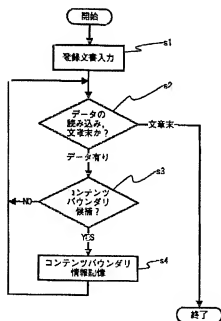
【図4】



【図6】



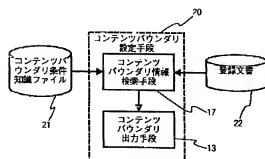
【図7】



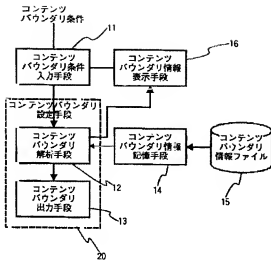
【図8】

コンテンツバウンダリ 識別番号	コンテンツバウンダリ 位置	コンテンツバウンダリ の種類	対応する コンテンツ バウンダリ	ネスト レベル
1	0	ファイル端	140	1
2	10	句点	1	2
3	20	句点	2	2
4	21	改行	1	2
13	43	リスト	24	2
14	49	改行	13	2
16	60	リスト項目	15	3
16	66	改行	16	3
17	67	リスト項目	17	3
24	89	リスト	13	2
25	96	改行	24	2
140	6405	ファイル端	1	1

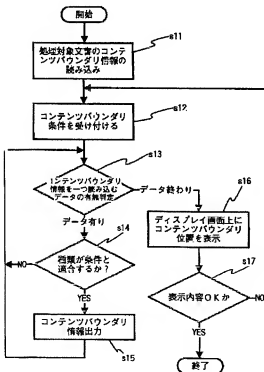
【図13】



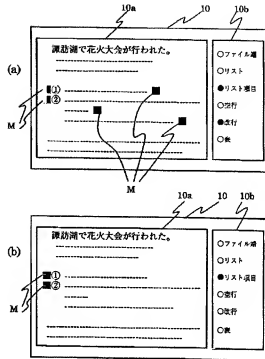
【图9】



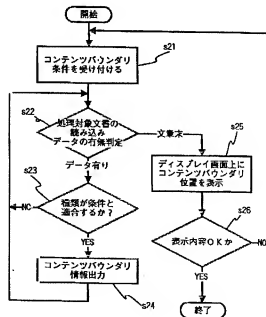
【图 1-1】



【图10】



【例 12】



フロントページの続き

(51)Int.Cl.⁶

識別番号

FI

G06F 15/401

310A

15/403

340A